Preprint
UCRL-JC-154129

Exhaustive Search for Fuzzy Gene Networks from Microarray Data

B.A. Sokhansanj, J.P. Fitch, J.N. Quong, A.A. Quong

This article was submitted to
25th Annual International Conference of the Institute for Electrical and Electronics Engineers Engineering in Medicine and Biology Society
Cancun, Mexico
September 17-21, 2003

July 7, 2003

Exhaustive Search for Fuzzy Gene Networks from Microarray Data

B. A. Sokhansansanj[1], J. P. Fitch[2], J. N. Quong[1], A. A. Quong[1]
[1]Chemistry & Materials Science Directorate, [2]Homeland Security Directorate, Lawrence Livermore National Laboratory, University of California, Livermore, CA

Abstract— **Recent technological advances in high-throughput data collection allow for the study of increasingly complex systems on the scale of the whole cellular genome and proteome. Gene network models are required to interpret large and complex data sets. Rationally designed system perturbations (e.g. gene knock-outs, metabolite removal, etc) can be used to iteratively refine hypothetical models, leading to a modeling-experiment cycle for high-throughput biological system analysis. We use fuzzy logic gene network models because they have greater resolution than Boolean logic models and do not require the precise parameter measurement needed for chemical kinetics-based modeling. The fuzzy gene network approach is tested by exhaustive search for network models describing cyclin gene interactions in yeast cell cycle microarray data, with preliminary success in recovering interactions predicted by previous biological knowledge and other analysis techniques. Our goal is to further develop this method in combination with experiments we are performing on bacterial regulatory networks.**

Keywords—**Gene networks, gene regulation, microarrays, simulation, fuzzy logic**

I. Introduction

Technological advances in DNA sequencing [1] have made it feasible to obtain the entire genetic sequence (genome) of an organism being studied by biologists. While the genomes of plants and animals are generally large (10^8-10^{10} bases, O(10^4) genes) and still take months and years to sequence, it is now possible to generate the draft genome sequence of a bacterium (10^6 bases, O(10^3) genes) in a matter of days or even hours. However, the sequence of genes only provides a "parts list" for the cell. Cell function arises from the regulatory pathways and networks of the genes and their protein products: how the parts are assembled and work together in response to environmental stimuli.

We are now in the "Post-Sequencing" era of biotechnology, characterized by engineering advances (reviewed in [1]) such as DNA chips and microarrays for mRNA transcript profiling, as well as protein profiling with mass spectroscopy and 2-D gel electrophoresis [2]. These technologies allow us to observe the activity of thousands of genes and proteins simultaneously, and we can use them to study the regulatory networks of the cell as an integrated unit. Furthermore, new genetic technologies, such as small interfering RNA (siRNA) for gene suppression facilitate high-throughput massively parallel perturbation of biological systems [3]. Given the complexity of the systems being analyzed and the size of data sets being generated, biologists need a modeling and simulation framework to optimally design experiments and interpret results.

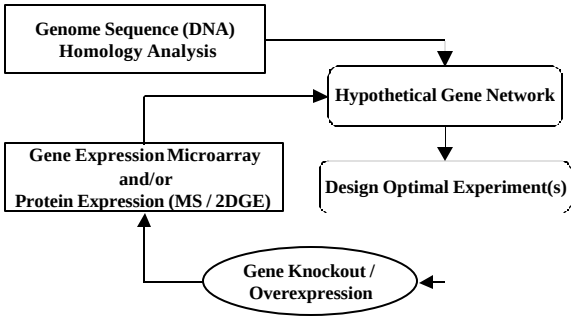

Fig. 1. Schematic of an integrated high-throughput modeling and experimental approach to analyzing genomic and proteomic regulation.

Computational Hypothesis Generation. Fig. 1 shows a schematic of an integrated experiment and modeling approach for high-throughput analysis of the genomic regulation mediating the response of a cell to a stimulus (e.g. temperature, host invasion, DNA damage, intercellular signaling molecule). Currently, the most convenient high throughput measurement technologies are DNA chips and microarrays for mRNA expression and 2D gel electrophoresis and mass spectrometry for protein expression. Future technologies which have already been demonstrated as proofs of concept include proteome-wide protein-protein hybridization [4] and genome-wide transcription factor hybridization [5].

III. Fuzzy Network Models

Many methods of modeling gene (and/or protein) networks based on expression data have been described in the literature. Boolean networks (e.g. [6]) are computational simple and thus suitable for handling both the complexity of biological networks and the challenge of generating and comparing multiple hypothetical networks as described in the above scheme. However, a Boolean model can not represent the true continuous nature of gene and protein levels, which are required to accurately represent the biological network. However, chemical kinetics based models (e.g. [7]) are both computationally complex and more importantly, sensitive to parameters that can not be accurately measured with inexpensive and high-throughput technologies. Statistical clustering methods are limited to describing correlation and anti-correlation between genes;

they can not represent complex functional relationships between multiple interacting genes.

Fuzzy logic [8] provides a mathematical framework that is compatible with poorly quantitative data. Furthermore, the language of fuzzy logic is consistent with the qualitative linguistic-graphical methods conventionally used to describe biological systems. The problem of rule set scalability is addressed by the union rule configuration (URC) developed by Combs and Andrew [9], which allows for linear growth in rule set complexity with both resolution and number of inputs at the cost of having to represent nonlinear relationships as "hidden layers". In the URC scheme, each input-output relationship is modeled by a single-antecedent fuzzy relation, and the final result is obtained by an aggregate fuzzy OR operation. Non-scalable conventional fuzzy logic has previously been used to analyze microarray data [10], and URC fuzzy logic has previously been used to model the *lac* operon of *E. coli* [11].

For our analysis, we use three fuzzy sets, LOW (or **1**), MED (**2**), and HIGH (**3**) to represent the magnitude of gene expression, as defined in Fig. 2. Experimental data is projected on to the interval [-1,+1]; currently this is done for Log 2 expression ratios by normalizing all data by the maximum value. For data that shows saturation characteristics, alternative nonlinear and piecewise linear projection functions may be considered. Defuzzification is performed using the centroid method [7], with point set definitions shown in Fig. 2.

As in the URC scheme, the rule for each input to a node is evaluated separately, with the final sum of the membership values in **1**-**3** taken across all rule evaluations and used for defuzzification to evaluate the output at the node. While inputs can be non-uniformly weighted, doing so comes at a significant cost in computational complexity. Under the three set scheme, there are 27 possible rules describing the effect of one gene on another gene. Thus, given a node (gene) with N input genes affecting it, there are 27^N possible rule combinations.

To further illustrate fuzzy gene networks, we assume a data set with three genes (G1, G2, G3) evaluated at three different times.

G1 = -1.0, 0, 1.0
G2 = -0.2, 0, 0.2
G3 = 0.6, 0, -0.6

Assuming that G3 is the node (output gene) and G1 and G2 are the inputs in our miniature regulatory network, only one combination of rules on G1 and G2 exactly fit the data:

G1:G3 = (**3 2 1**)
G2:G3 = (**3 2 1**)

These rules can be read as

If G1 is Low (**1**) then G3 is High (**3**)
If G1 is Med (**2**) then G3 is Med (**2**)
If G1 is High (**3**) then G3 is Low (**1**) ... etc.

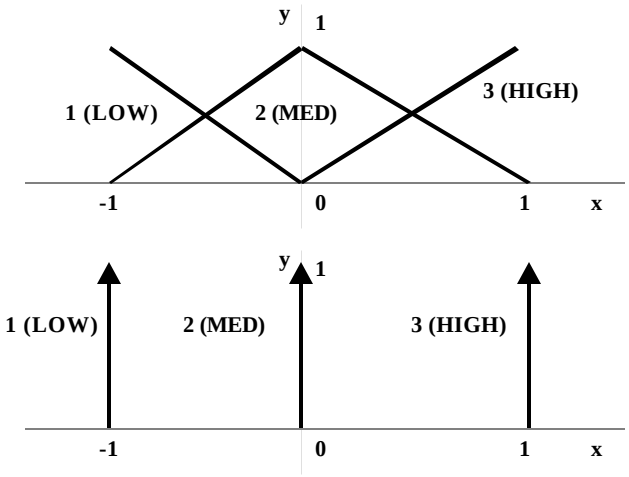

Fig. 2. Fuzzification and defuzzification with experimental data on the x-axis and evaluated membership functions on the y-axis.

For real data, in general no rule combination will be an exact fit, and given some tolerance there will be multiple possible rule combinations, representing plausible hypothetical gene network models. We can use the fit error to rank these models, and use clustering schemes to find common structures in the models to help design experiments that can optimally differentiate between alternate models. (While under the scheme of Fig. 2 there is no inherent error in the fuzzification and defuzzification process, other schemes may result in a finite error, representing a minimum "best fit" tolerance.)

IV. Yeast Cyclin Array Data Analysis

As a proof of concept, we have used exhaustive search to generate fuzzy gene networks based on microarray data obtained for the yeast cell cycle time series by Spellman, *et al*. [12], a data set frequently used by researchers validating analysis methods. We focus on the network of interactions among yeast cyclin proteins, for which a mathematical model has developed [13]. Notably, gene networks obtained from microarray data represent both direct interactions between transcription factors and the genes they regulate, as well as indirect interactions mediated through post-translational modifications, metabolite fluxes, and protein-protein interactions.

We focus in particular on seven well-characterized yeast cyclins, coded by the genes CLN3, SWI5, HCT1, CDC20, SIC1, CLB2, and CLB5; generating hypothetical fuzzy networks based on assuming each as a node and the others as inputs. In one case, we exclude SWI5, which is known to be a key transcriptional factor that is itself transcriptionally regulated (and thus a key element of any cyclin gene network model), and in the other HCT1, which acts and is regulated post-translationally and thus only an indirect

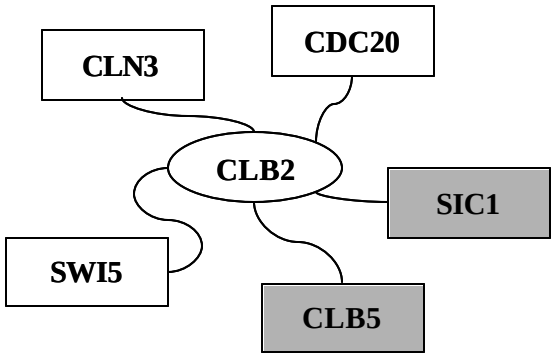

Fig. 3. Known relationships (positive, white boxes; negative, gray boxes) between yeast cyclins and the CLB2 gene. Includes both direct transcriptional relationships (CLB2 inhibits SIC1) and indirect relationships, i.e. expression during adjacent cell cycles (e.g. CLN2 is G1/S specific, CLB2 is G2/M specific).

component of a gene network.

To illustrate the analysis, we will focus on the fuzzy gene network at the CLB2 node assuming CLN3, SWI5, CDC20, SIC1, CLB5 as inputs. Existing biological knowledge for CLB2 regulation is shown in Fig. 3, based on direct and indirect relationships (as described in [13]). The best-fit (minimum RSS error) fuzzy rules to the Spellman data set (for cdc15 cell synchronization) were found by exhaustive search and are shown in Fig. 4. As these rules show, even though there are several rules that fit almost equally well, there is a clear pattern that is qualitatively consistent with the relationships shown in Fig. 3. We are currently developing methods to summarize and visualize the results of exhaustive network search, for brevity here we will interpret "rule histograms": the number of each of the 27 possible rules that one of the input genes can take in each rule combination. For example, in Fig. 4, SIC takes the rule (**3 1 1**) 13/14 times, consequently it is a significant rule.

Fig. 5 shows the predicted CLB2 time series of the best fit rule (first line of Fig. 4) given the yeast cell cycle time series synchronized by cdc15 addition (used for network fitting), as well as data generated for cdc28, and alpha factor addition (see [12] for a detailed description of experimental protocols). The rule set has a good qualitative fit with the cdc28 and alpha factor time series (noting that the alpha set has several missing values and the magnitude of the response is highly sensitive to experimental conditions). Also, it should be noted that in fitting the rule set we have assumed the contribution of all input genes are equally weighted.

Err	CLN3	SWI5	CDC20	SIC1	CLB5
1.773	1 1 3	1 3 3	1 2 3	3 1 1	3 3 1
1.774	1 1 3	1 2 3	1 3 3	3 1 1	3 3 1
1.777	1 2 3	1 2 3	1 2 3	3 1 1	3 3 1
1.809	1 1 2	1 2 3	1 3 3	3 1 1	3 3 1
1.810	1 2 2	1 2 3	1 2 3	3 1 1	3 3 1
1.811	1 1 2	1 3 3	1 2 3	3 1 1	3 3 1
1.824	1 2 3	1 2 3	1 3 3	3 1 1	3 2 1
1.826	1 1 3	1 2 3	1 2 3	3 2 1	3 3 1
1.826	1 2 3	1 1 3	1 3 3	3 1 1	3 3 1
1.835	1 2 3	1 3 3	1 1 3	3 1 1	3 3 1
1.846	2 1 3	1 3 3	1 2 3	3 1 1	3 3 1
1.848	1 2 3	1 3 3	1 2 3	3 1 1	3 2 1

Fig. 4. Best fit fuzzy network rules for CLB2. The first line of the rule table indicates that the CLN3:CLB2 rule is (1 1 3), the SWI5:CLB2 rule is (1 3 3), etc. The CLN3:CLB2 rule is read in English as "If CLN3 is Low (**1**) Then CLB2 is Low (**1**); If CLN3 is Medium (**2**) Then CLB 2 is Low (**1**); If CLN3 is High (**3**) Then CLB2 is High (**3**).

We have compared the results of our analysis with a supervised learning scheme described in [14]. In the analysis of [14], only two states of expression were considered, "under" and "over", with thresholds inferred from the data. While both analyses generate predictions consistent with known biology, fuzzy gene network analysis is more sensitive to small changes in transcription level, includes more details of functional relationships, and consequently fuzzy analysis can pose potential alternative hypotheses that may be consistent with the nonlethality of certain cyclin mutations (e.g. CLB5-6). For example, in [14] no rules were found involving HCT1, which encodes a protein responsible for degrading Clb2. Hct1 acts posttranslationally on Clb2, and shows only small (but detectable) cycling expression levels that can be deemed too "insignificant" by statistical modeling techniques. However, the network search revealed significant rules for CLN3:HCT1 (**1 3 3**) corresponding to indirect relationships between Hct1 and Cln3 (Cln3 activates SBF and Hct1 degrades Clb2, an SBF inhibitor) responsible for driving the cell cycle.

V. Conclusion

We have shown preliminary success in analyzing simulated and actual data sets (i.e. yeast cell cycle data). As shown in Fig. 5, linear combinatorial fuzzy networks (unlike, e.g. Boolean networks) have sufficient resolution to accurately reproduce complex time series data sets, without using continuous parameters. In addition, we can model key biological details of yeast cyclin interactions. We are further developing our method in close integration with gene and protein expression experiments performed in our own laboratory. The most significant drawback of the fuzzy gene network modeling method is the computational complexity associated with exhaustive search; particularly when there is no connectivity network inferred from DNA sequence comparison and no simplifying structure can be imposed on the problem. Thus, we are currently exploring heuristic methods for combinatorial optimization (i.e. genetic algorithms), as well as integrating Bayes net methods with fuzzy logic representation of functional relationships. It is important to note that the gene network inference problem is not a classical inverse problem with a static data set.

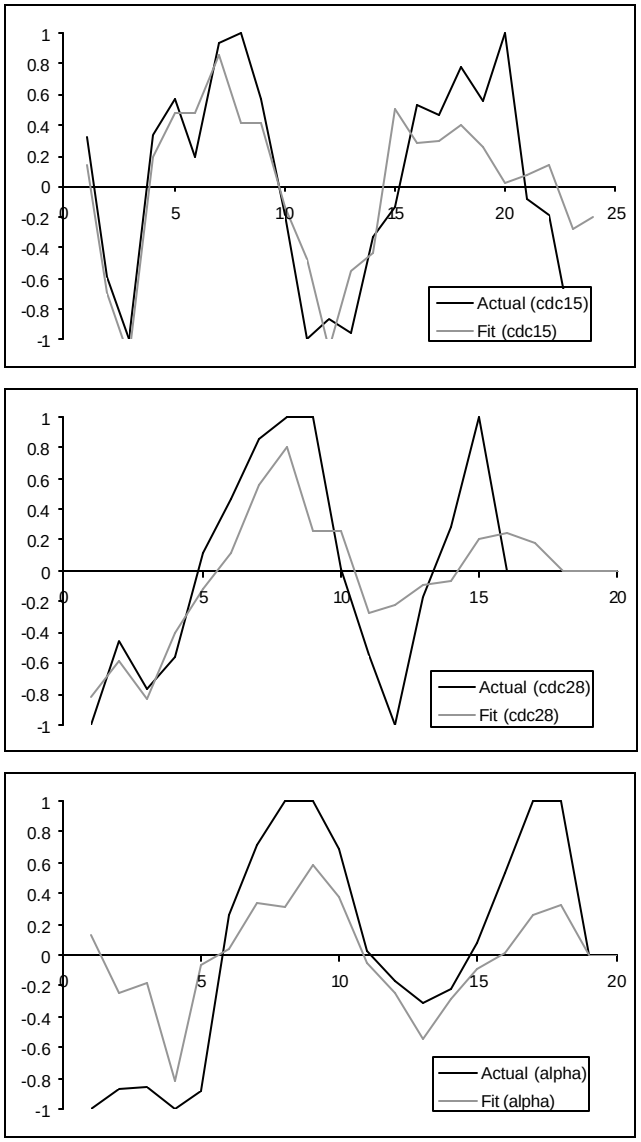

Fig. 5. Comparison of experimental ([10]) and predicted CLB2 yeast cell cycle time series assuming different synchronization methods.

Rather, biological system analysis must be treated as a dynamic reverse engineering problem, in which searching for the solution is continually assisted by optimal acquisition of new experimental data. As such, the methods developed for model-assisted experimental design in gene networks may have significant application for other network and related reverse engineering problems.

ACKNOWLEDGMENT

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48. This research is funded by the Laboratory Directed Research and Development (LDRD) Program at Lawrence Livermore National Laboratory (LLNL). The LDRD Program is mandated by Congress to fund director-initiated, long-term research and development (R&D) projects in support of the DOE and national laboratories mission areas. The Director's Office LDRD Program at LLNL funds creative and innovative R&D to ensure the scientific vitality of the Laboratory in mission-related scientific disciplines.

REFERENCES

[1] J. P. Fitch and Sokhansanj, B., "Genomic engineering: moving beyond DNA sequence to function," *Proc. IEEE*, vol. 88, pp. 1949-1971, Dec. 2000.

[2] Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L., and Aebersold, R., " Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae," *Mol. Cell Proteomics*, vol. 1, pp. 323-333, Apr. 2002.

[3] J. Pothof, G. van Haaften, K. Thijssen, R. S. kamath, A. G. Fraser, J. Ahringer, R. H. Plasterk and M. Tijsterman, "Identification of genes that protect the C. elegans genome against mutations by genome-wide RNAi," *Genes Dev.*, vol. 17, pp. 443-448, Feb. 15, 2003.

[4] Bader, G. D. and C. W. Hogue, "Analyzing yeast protein -protein interaction data obtained from different sources," *Nat. Biotechnol.*, vol. 20, pp. 991-997, Oct. 2002.

[5] Lee, T. I., *et al*., "Transcriptional regulatory networks in Saccharomyces cerevisiae," *Science*, vol. 298, pp. 763-764, Oct. 25, 2002.

[6] S. Liang, Fuhrman, S., and Somogyi, R. "REVEAL, a general reverse engineering algorithm for inference of genetic network architectures," *Pacific Symposium on Biocomputing*, vol. 3, pp. 18-29, 2000. [Online]. http:// www-smi.stanford.edu/projects/helix/psb98/

[7] D. Endy, You, L, Yin, J., and Molineux, I. J., "Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes," *Proc. Natl. Acad. Sci. USA*, vol. 97, pp. 5375-5380, May 9, 2000.

[8] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338-352, 1965.

[9] W. E. Combs, and Andrews, J. E., "Combinatorial rule explosion eliminated by a fuzzy rule configuration," *IEEE Trans. Fuzzy Syst.*, vol. 6, pp. 1-11, Feb. 1998.

[10] Woolf, P. J., and Wang, Y., "A fuzzy logic approach to analyzing gene expression data," *Physiol. Genomics*, vol. 3, pp. 9-15, Jan.-Feb., 2000.

[11] Sokhansanj, B. A., Garnham, J. B., Fitch, J. P., "Interpreting data from microarray experiments to build models of microbial genetic regulation networks," SPIE Photonics West: Biomedical Optics and Applications (BiOS 2002), San Jose, CA, Jan. 19-25, 2002; Proc. SPIE Functional Monitoring and Drug-Tissue Interaction, Vol. 4623, pp. 27-37.

[12] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B., "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Molec. Biol. Cell*, vol. 9, pp. 3273-3297, Dec. 1998.

[13] Chen, K. C., Csikasz-Nagy, A., Gyorffy, B., Val, J., Novak, B. and Tyson, J. J., "Kinetic analysis of a molecular model of the budding yeast cell cycle," *Molec. Biol. Cell*, vol. 11, pp. 369-391, Jan. 2000.

[14] Soinov, L. A., Krestyaninova, M. and Brazma, A., "Towards reconstruction of gene networks from expression data by supervised learning," *Genome Biology*, vol. 4, R6, Jan. 6, 2003, available online http://genomebiology.com/2003/4/1/R6.